Scalable Homology Search Tools

Singh, A.K.
Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA, USA

The growth in biological information has spurred increased interest in scalable similarity comparison of different biological data types: sequences, structures, expression arrays, and pathways. We have developed new algorithms for this task. In the case of DNA and protein sequences, we first transform the database sequences into a vector space and store the summaries in an index structure. Given a query sequence, we identify possible interesting similar regions using the index structure; these regions can then be explored through a more intensive search tool such as BLAST. These vector-space transformations also have the potential for speeding up multiple alignments. In the case of protein structures, we summarize the local neighborhoods of each protein in a given database of protein structures and store the summaries in an index structure. Given a query structure, similar regions can be identified using the index structure and graph analysis. This search technique has the potential for quickly identifying motifs. In the case of pathways, we have developed a tool that can measure the distance between pathways given the similarity between the enzymes (or other elementary objects) and their interconnections. These distances can be used for analysis of similar pathways, such as the construction of phylogenetic trees based on a specific metabolism function on a group of organisms.